DIAGNOSIS OF BREAST CANCER USING CASE-BASED REASONING

NURRAMLAH BINTI ABU NASIR

A Thesis submitted in partial fulfilment of the requirement for the awarded of the Bachelor of Computer Science and Software Engineering

Faculty of Computer Systems & Software Engineering
University Malaysia Pahang

Jun, 2012

STUDENT'S DECLARATION

"I hereby declare that this report and entitled 'Diagnosis of Breast Cancer Using Case-Based Reasoning' is the result of my research. The report has not been accepted for my bachelor and it's not concurrently submitted in candidate of my bachelor"

Signature :…….

Student : NURRAMLAH BINTI ABU NASIR

Date :……

SUPERVISOR'S DECLARATION

"I hereby declare that I read this thesis and in my opinion is sufficient in terms of scope and quality for the award of Bachelor of Computer Science and Software Engineering"

Signature :……..

Supervisor : MOHD HAFIZ BIN MOHD HASSIN

Date :……..

DEDICATION

Special dedication to my family members especially to my beloved father and mother
(Encik Abu Nasir Bin Misdi and Puan Masamah Binti Saliman)
who always give me encouragement in my life, my study and to finish
my Undergraduate Project 2.
To my Supervisor

Mr.Mohd Hafiz Bin Mohd Hassin

To my Academic Advisor

Mr. Ng Choon Ching

To all FSKKP's lecturers

To all my classmate, 3BCS

And all my friends outhere

Thank You for your supporting and teaching

Thank You for everything that gave during my studies and the knowledge that we shared together.

Lastly, to my fiancée

Ahmad Affendi Bin Mat Arifin

Thanks for giving me your supports deeply inside my heart and soul and inspired me along my life and my study.

THANK YOU SO MUCH

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and Most Merciful, first of all, I would like to express gratitude to Allah for all His Divine Guidance. Praise be to Allah, by the end of this course, I can complete my project successfully with the inspiration, grace and guidance has been given by Him.

Firstly, I would like to express my sincere appreciation to my project supervisor, Mr. Mohd Hafiz Bin Mohd Hassin for his guidance and supports during this project. He never felt tired in giving me and my colleagues motivation, advice and constant throughout the duration to finishing this PSM. Without the guidance of such a strong and cohesive than this which I will probably be able to complete my project and my thesis. Besides that, thanks a lot to my parents and my fiancée for their supporting, understanding and love along my life and never stop in giving me inspire and support.

I'm very grateful towards to all the lecturers in Faculty of Computer Systems & Software Engineering (FSKKP), Universiti Malaysia Pahang (UMP) who have directly or indirectly influential and give supportive for this PSM and also lend me their hands in solving problems during the development of the application.

I also appreciation to my friends that also help in providing guidance and ideas in an effort to me. Lastly, I would like to thank everybody who was involved to the successful of my project. May Allah bless you all. Thank you very much.

ABSTRACT

The objective for this thesis is to develop an intelligent decision support application for diagnosis of breast cancer using Case-Based Reasoning (CBR) algorithm for predict the class of cancer for patients. The number of expertises in the medical domain about the breast cancer is limited. Many patients have to wait too long to get their result from the check-up. The experience medical staffs are decreasing in number. When they retired, the new staffs will be replacing their places. So they have to learn many things related to their work. The application is very useful in the management of the problem and aids the inexperience physicians to check their diagnosis. It is to help the expert doctors or medical staffs in their breast cancer diagnosis. The methodology used in the application is Rapid Application Development (RAD) because it promotes the accuracy application development and delivery and reduced the cycle time. The application used the 100 data of Wisconsin Breast Cancer dataset for evaluating the CBR algorithm. This dataset is retrieved from UCI Machine Learning. The data used in this application consists of 9 attributes where the result of each case will be classified either non-cancerous (*benign)* or cancerous (*malignant)* group.

ABSTRAK

Objektif tesis ini adalah untuk membangunkan satu aplikasi sokongan keputusan pintar untuk mendiagnosis kanser payudara menggunakan algoritma Penaakulan Berasaskan Kes (CBR) untuk meramalkan kelas kanser bagi para pesakit. Bilangan pakar dalam bidang perubatan adalah terhad. Ramai pesakit yang perlu menunggu lama untuk mendapatkan keputusan pemeriksaan mereka. Kakitangan perubatan yang berpengalaman juga berkurang. Apabila mereka bersara, kakitangan baru akan menggantikan tempat mereka. Jadi, mereka perlu mempelajari banyak perkara berkaitan dengan pekerjaan tersebut. Aplikasi ini sangat berguna dalam masalah pengurusan dan membantu kakitangan yang tidak berpengalaman membuat pemeriksaan diagnosis mereka. Aplikasi ini adalah untuk membantu doktor-doktor pakar ataupun kakitangan perubatan dalam mendiagnosis kanser payudara mereka. Metodologi yang digunakan dalam aplikasi ini ialah Pembangunan Aplikasi Pantas (RAD) kerana ia adalah pemangkin pembangunan aplikasi dan penghantaran ketepatan dan mengurangkan masa kitaran. Projek ini menggunakan 100 data dari set data Kanser Payudara Wisconsin untuk menilai algoritma CBR. Data set ini diambil daripada mesin belajar UCI. Data yang digunakan dalam projek ini terdiri daripada 9 sifat yang mana hasil bagi setiap kes akan diklasifikasi sama ada kumpulan bukan kanser (*benign*) ataupun kanser (*malignant*).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANCRONYM	MEANING
AI	Artificial Intelligence
BCDA	Breast Cancer Diagnosis Application
CBR	Case Based Reasoning
EHR	Electronic Health Records
IDE	Integrated Development Environment
RAD	Rapid Application Development
RAM	Random Access Memory
SVM	Support Vector Machine

LIST OF APPENDICES

CHAPTER I

INTRODUCTION

1.0 Introduction

This chapter briefly describes the diagnosis of breast cancer using case-based reasoning (CBR) for predict the class of cancer for patients that have been developed. This chapter consists of six sections: The first section describes the background of the project. The second section describes the problem statement and motivation of the project. The third section describes the objectives for the project. The fourth section describes the scopes for the project and section five briefly explains about the project contribution. Finally in section six, the thesis organization is described.

1.1 Background

The term of *breast cancer* is referred to a *malignant* tumour that has developed from cells in the breast. Mostly, it is found in women but men also can get breast cancer even it is rare. It is affecting about 10% of all women at some stages in their life (cancer.org, 2011). From American Cancer Society, breast cancer is the second reason of death for women in the United States. Factors that may cause the breast cancer are age, family history, race and many more. A tumour can be classified as *benign* or *malignant*.

CBR is an approach for solving problems based on solution of similar past cases. The purpose of CBR is to provide the decision maker with the ability to utilize the specific knowledge of previously experienced, concrete problem situation or specific patients' cases. This approach for developing knowledge-based medical decision support applications based on the new technology of CBR. It is not used only in medical domain but also in the financial, agricultural, management and many more domains (Investopedia.com, 2011). There are some different papers present the technique for diagnosis of breast cancer such as Neutral Network, Expert Application, Boosted Decision Tree and their ensemble combination. By using the CBR technique, it can improve the solving problem performance through reuse and makes use of existing data. It also can reduce the knowledge acquisition efforts and require less maintenance effort.

1.2 Problem Statement

The number of specialists and expertises in the medical domain about the breast cancer is limited. The patients have to make appointments with them before doing the medical check-up. So many patients have to wait too long to get their result from the check up. CBR will help the doctors to make their works easy and provide the quick and correct medical reports to their patients.

Besides that, from day to day, the experience medical staffs are decreasing in number. When they retired, the new staffs will be replacing their places. They are inexperience staff compared to the old one. So they have to learn many things regarding to the related works. The application is very useful in the management of the problem and aids the inexperience physicians to check their diagnosis.

1.3 Objectives

The objectives of the application are:

1) To develop an intelligent decision support application for diagnosis the breast cancer in order to classify the class of tumour either *benign* or *malignant* group.
2) To apply the CBR algorithm in the breast cancer diagnosis application.

1.4 Scope

This artificial intelligence application uses the CBR algorithm to solve the problem and used the breast cancer Wisconsin dataset that retrieved from UCI Machine Learning. This application is based on stand-alone application and focus on classifying the class of tumour as either *benign* or *malignan*t group. The user for the application is expert doctors and medical staffs in order to help them diagnose the breast cancer.

1.5 Project Contribution

CBR methods help to compensate for lack of experience of young medical staffs. The inexperience staffs need the guide from the experience staffs to improve their skill in handling the diagnosis. It also to reduce the time required to come to a decision particularly in an emergency case.

1.6 Thesis Organization

This application is about CBR for diagnosis breast cancer and consists of the following chapters: Chapter 1 is introduction. In this chapter, it provides background information about the application which normally includes problem statement, objectives and scope. Chapter 2 consists of literature review. It is a critical and in depth evaluation of previous research of the CBR technique in diagnosis breast cancer in order to prove the statement by the citation. In Chapter 3, methodology is a guideline for solving a problem, with specific components such as phases, tasks, methods, techniques and tools using in the project. In Chapter 4, the implementation of the research project is

presented. For Chapter 5, result, discussion part, present the result of the project and discuss the outcome of the project. In Chapter 6, conclusion and recommendation part, the researcher makes the conclusion and suggests some recommendation in order improve the project in future. This chapter will briefly summarize the overall project.

1.7 Summary

This chapter explain the overview of the overall of the thesis. The aim of this application is to help doctors classify the class of tumour as either *benign* or *malignant* group. While the objectives are to develop an intelligent decision support application for diagnosis the breast cancer in order to classify the class of tumour either *benign* or *malignant* group and to apply the CBR algorithm in the breast cancer diagnosis application. It is hoped that the CBR algorithm can classify the patients' case either *benign* group or *malignant* group and achieved all the objectives stated. The literature review will be discussed about the previous works done by others researchers in next chapter.

CHAPTER II

LITERATURE REVIEW

2.0 Introduction

This chapter briefly describes the review on diagnosis of breast cancer using case-based reasoning. In this chapter, three sections are comprised: The first section briefly explain the background of breast cancer and breast cancer dataset. The second section describes the review on artificial intelligence techniques includes neural network, genetic algorithm, support vector machine (SVM), fuzzy logic and CBR. The third section describes about the proposed technique used for the project which is CBR. It also includes the CBR cycle, applications of CBR and advantages of CBR.

2.1 Breast Cancer

Breast cancer is one of the common cancers in the world. It is a tumour that has developed from the cell of the breast. It occurs in men and women even though the breast cancer in men is rare case. Breast cancer is the leading cause of death among women between 40 and 55 years old and the second overall cause of death among women (exceeded only by lung cancer) (Imaginis.com, 2011). According to the World Health Organization, more than 1.2 million women will be diagnosed with breast cancer each year worldwide. White women have a higher incidence of breast cancer than African American women beginning at age 45 (cancer.org, 2011). Actually not all tumours are cancer. Tumours can be non-cancerous (*benign*) or cancerous (*malignant*).

2.1.1 Breast Cancer dataset

The breast cancer dataset used was collected by Dr. William H.Wolberg in year 1989 to 1991 at the University of Wisconsin-Madison Hospitals (Mangasarian and Wolberg, 1990). This dataset can be retrieved from UCI Machine Learning Repository. The dataset contains 699 samples with 16 samples have attributes with missing values and 683 samples have complete data. Each sample record has nine attributes and graded on an interval scale from 1 – 10. They are queue in order of ordinal data type (ordered set). The class attribute is come with two states; 2 for *benign* and 4 for *malignant*.

1. Number of attributes: 11 (id numbers, 9 integer-valued attributes and 1 class attribute)

2. Attributes information: (name of attributes and the type of value domain)

Table 1.1 Attributes information and type of value domain

Sample code number	id number
Clump Thickness	Ordinal (integer value in range [1-10])
Uniformity of Cell Size	Ordinal (integer value in range [1-10])
Uniformity of Cell Shape	Ordinal (integer value in range [1-10])
Marginal Adhesion	Ordinal (integer value in range [1-10])
Single Epithelial Cell Size	Ordinal (integer value in range [1-10])
Bare Nuclei	Ordinal (integer value in range [1-10])
Bland Chromatin	Ordinal (integer value in range [1-10])
Normal Nucleoli	Ordinal (integer value in range [1-10])
Mitoses	Ordinal (integer value in range [1-10])
Class	2 for *benign*, 4 for *malignant*

The output from attributes is to predict the class of cancer. This is known as the supervised learning. This learning will be explained in the next topic.

2.2 Artificial Intelligence Techniques

Artificial Intelligence (AI) is the part of computer science concerned with designing intelligent computer applications, that is, applications that exhibit characteristics we associate with intelligence in human behaviour – understanding language, learning, reasoning, solving problems and so on (Roukis.et.al, 1990). Any situation in which both the inputs and outputs of a component can be perceived is called supervised learning. It is a technique for creating a function for a training data. The training data consists of pairs of input objects (a vector of characteristics) and desired output. Many cases in AI techniques have been used to support learning techniques and improve problem-solving strategies. In the literature papers, the diagnosis of breast cancer can be applied in many AI techniques. AI consists of many sub-fields using a variety of techniques such as Neural Network, Genetics Algorithm, Support Vendor Machine (SVM), Fuzzy Logic and CBR.

2.2.1 Neural Network

Neural network is a series of algorithm that attempt to identify underlying relationship in a set of data by using a process that mimics the way the human brain operates. Neural network have the ability to adapt to changing input so that the network produces the best possible result without the need to redesign the output criteria (Investopedia.com, 2011). Evolving neural networks for detecting breast cancer is applied (Fogel.et.al, 1995). Artificial Neural Network is a branch of computational intelligence that employs a variety of optimization tool to learn from past experiences and use that prior training to classify new data, identify new patterns or predict (Jangel

et.al. 2010). The application of neural networks is in pronunciation, handwritten character recognition, mobile computing and so on. A Neural network learns and does not need to be reprogrammed and can be implemented in any application.

The advantage using neural network is a neural network learns application behaviour by using application input-output data. Neural networks have good generalization capabilities. The learning and generalization capabilities of neural nets enable it to more effectively address nonlinear, time variant problems, even under noisy conditions. Thus, neural nets-can solve many problems that are either unsolved or inefficiently solved by existing techniques, including fuzzy logic. Finally, neural networks can develop solutions to meet a pre-specified accuracy. It is difficult, if not impossible, to determine the proper size and structure of a neural networks to solve a given problem. Neural network also do not scale well. Manipulating learning parameters for learning and convergence becomes increasingly difficult. Artificial neural network are still far away from biological neural networks, but what we know today about artificial neural networks is sufficient to solve many problems that were previously unsolvable or inefficiently solvable at best.

2.2.2 Genetics Algorithm

Genetic algorithm is a popular technique used for searching large solution spaces. It may help the physician in prediction several cases of health. An advantage of genetic algorithm is these applications go through an iterative process to produce an optimal solution (Wikipedia.com, 2011). The fitness function determines the good solutions and the solutions that can be eliminated. A disadvantage is the lack of

transparency in the reasoning involved for the decision support applications making it undesirable for physicians. Genetic algorithm has proved to be useful in the diagnosis of female urinary incontinence. It always appears "premature" convergence or convergence slow shortcomings in the actual application (WangLing, 2001). Genetic algorithm will improve generalization using classification accuracy as the fitness function. The genetic algorithm also can minimize cost and maximizing accuracy.

2.2.3 Support Vector Machine (SVM)

SVM is a learning machine used as a tool for data classification, function approximation, etc, due to its generalization ability and has found success in many applications (Cortes and Vapnik, 1995). The SVM proposed by Vapnik (1995) has been studied extensively for classification, regression and density estimation. Feature of SVM is that it minimizes and upper bound of generalization error through maximizing the margin between separating hyper plane and dataset. SVM has an extra advantage of automatic model selection in the sense that both the optimal number and locations of the basis functions are automatically obtained during training. The SVM classifiers showed a great performance since it maps the features to a higher dimensional space. The algorithm of SVM is able to create a complex decision boundary between two classes with good classification ability (Elsayad, 2010).

2.2.4 Fuzzy Logic

Fuzzy logic is a mathematical logic that attempts to solve problems by assigning values to an imprecise spectrum of data in order to arrive at the most accurate conclusion possible (Investopedia.com, 2011) .Fuzzy logic is designed to solve problems in the same way that humans do: by considering all available information and making the best possible decision given the input. In medical domain, the fuzzy logic in computer-aided breast cancer diagnosis; analysis of lobulation by Kovalerchuk et.al. (1997) is applied.

The advantages using fuzzy logic converts complex problems into simpler problems using approximate reasoning. The application is described by fuzzy rules and membership functions using human type language and linguistic variables. Thus, one can effectively use his/her knowledge to describe the application's behaviour. A fuzzy logic description can effectively model the uncertainty and nonlinearity of a application. It is extremely difficult, if not impossible, to develop a mathematical model of a complex application to reflect nonlinearity, uncertainty, and variation over time. Fuzzy logic avoids the complex mathematical modeling. Fuzzy logic is easy to implement using both software on existing microprocessors or dedicated hardware. Fuzzy logic based solutions are cost effective for a wide range of applications (such as home appliances) when compared to traditional methods. But the Fuzzy logic uses heuristic algorithms for defuzzification, rule evaluation, and antecedent processing. Heuristic algorithms can cause problems mainly because heuristics do not guarantee satisfactory solutions that operate under all possible conditions.

2.2.5 Case Based Reasoning

CBR has been applied in many domains to solve variety of problem such as medical diagnosis, aircraft maintenance, credit card risk assessment and so on (ai-cbr.org, 2011). Research conducted in lab shows both novice and experienced car mechanics use their own experiences and those of others to help them generate hypotheses about what is wrong with a car, recognize problem and remember how to test for different diagnoses (Lancaster and Kolodner, 1988).

CBR can solve a few problems. A representation form for cases has been determined and an appropriate retrieval algorithm has been selected. The objective and subjective knowledge can be clearly separated. So they can be used together in one application. The problem of updating the changeable subjective knowledge can partly be solved by incrementally incorporating new up-to-date case. This algorithm gives a hierarchical structure used to intelligently index the cases, thus avoid the expensive linear search in large libraries of cases and significantly increasing recall accuracy. Users can get the accuracy and efficiency of the application without the problem of brittleness and maintenance.

2.3 The proposed technique

2.3.1 Case Based Reasoning

CBR is a methodology which originated in artificial intelligence community as an effective alternative to traditional rule-based applications for generating intelligent decisions in weak-theory application domains (Deng, 1996). CBR is an approach for solving problems based on solution of similar past cases (Kolodner, 1993). It also can be utilized to solve a new problem by remembering a previous similar situation by reusing information and knowledge of that situation (Aamodt and Plaza, 1994). It provides the decision maker with an ability to utilize the specific knowledge of previously experienced, concrete problems situation or specific patients' cases. CBR comprises four steps: retrieve a case that is the most similar to the new input problem, reuse and revise the solution of the retrieved case to generate a solution for the new problem, and retain the new input problem and its solution as a new case in the case base. Central tasks that all CBR methods have to deal with are identifying the current problem situation, find a past case similar to the new one, use the case to suggest a solution to the current problem, evaluate the proposed solution and update the application by learning from this experience (Aamodt and Plaza, 1994).The retrieval process involves the tasks of situation assessment, initial match and final selection.

2.3.2 The CBR cycle

The CBR cycle can be generally describe by the following four processes:

1) **Retrieve** the most similar case or cases
2) **Reuse** the information and knowledge in that case to solve the problem
3) **Revise** the proposed solution
4) **Retain** the parts of this experience likely to be useful for future problem solving

These features can be information such as patient information, factors that cause the cancer and etc. This information is then stored in a database for retrieval. Retrieve is the method of searching the database to find similarities between the cases based on the indexed information. The retrieved case combined with the new case through reuse into the solved case. The suggested solution for the initial problem. Adaptation is the changing of the retrieved information to best fit the new problem. Revise is the verifying step of the fitness of the proposed solution and determining if the proposed solution is success or the process needs to be started over again. The final step is retain which is the useful experience retained for future reuse and update the case based by a new solved case or by modification of some existing cases.

2.3.3 Applications of CBR

CBR becomes a successful technique for knowledge- based applications in many domains. CBR applications have been used in the medical domain for purpose of diagnostic, classification, tutoring and planning (such as therapy support) (Nilsson and Sollenborn, 2004). In medication, the knowledge of experts does not only consist of rules, but of a mixture of textbook knowledge and experience. The GS.52 application by

Gierl and Stengel-Rutkowski (1994) is an example that is in practical use as a real-life application. It uses CBR to address the domain of dismorphic syndromes that used in the children's hospital of the University of Munich for many years. Such a syndrome means a non-random combination of different disorder. The physician selects a new or an existing syndrome and typical cases for the syndrome. Subsequently, GS.52 determines the relevant features and their relative frequency. Individual knowledge processing equates to a CBR approach that employs collections of patients' cases. Such collections are focus of research on Electronic Health Records (EHR). EHR contain a wealth of data that could be used to support case-based decision. If EHR are to be used in a CBR context, the issues pertinent to the design of case-bases automatically become pertinent to the EHR design and the CBR paradigm becomes important to Medical Informatics (Pantazi.et.al, 2004).

CBR also can be applied in solving the examination timetable problem. Examination timetable is very important for students and teachers. From the timetable, they will know about the duration, the name of subjects and the place for examination. By using that, the timetable will be not overlapping or clash with other subjects. The problem is somewhere analogous to the problem of colouring an undirected graph in which a vertex correspond to an examination and an edge links two examination if they have one or more students in common (Welsh and Powell, 1967). In the domain of scheduling, some researchers have resorted to CBR to enhance the flexibility and robustness of scheduling. The CBR application in scheduling can be classified into three categories; algorithm reuse, operator reuse and solution reuse. A CBR framework is designed to choose an appropriate algorithm for a given production scheduling problem (Schmidt, 1998).

CBR approach is used to obtain the treatment trains for treating the wastewater. The research has been done by Krovvidy and Wee (1993) in treatment trains. A treatment train is a sequence of individual unit processes where the effluent of one process becomes the influent to the next process. The process of obtaining the optimal treatment trains using a heuristic search function involves a large search space. This search effort can be reduced if using some of the old treatment train already existing in the case base. CBR application addresses the memory problem by remembering a selected set of cases. The performance is found to be slightly better when a new case is stored only when it cannot be correctly classified by the instances currently in memory.

2.3.4 Advantages of using CBR

The CBR have their own advantages same like other algorithms. According to Pal and Shiu (2004), the advantages of CBR are as follow. First, CBR is learning over time. CBR used the past problem encountered to create the solution. From the past, it captures and indexes its mistakes. The past experiences provide warning for the reasoner so that they can avoid those past failures. As the cases are added into the case base, it can be used to help in solving future problems. Learning can be achieved regardless of success or failure since both failed and successful cases are stored in the database. Second, case-based reasoning provide a means of explanation. Problem solving CBR is using a previous method to help determine solutions to new problems. By describing the successful past cases using the similar new cases, the CBR can explain the solution for the future to user. Citing actual cases make the explanation more useful. Third, case-based reasoning also avoid repeating all the steps that need to be taken to arrive at a solution. CBR can reuse the solution without going into the sequences of tedious

question for new problem as the rule-based reasoning does. It solves problems by adapting old solutions without any need to derive answers from scratch each time.

2.4 Summary

The number of expertises in the medical domain about breast cancer is limited. Many patients have to wait too long for getting the check up result. The experience medical staffs are decreasing in number due to retire, the new staffs will replacing their places. They have to learn more about their works. This application is very useful in the management application and aids the inexperienced physicians to check their diagnosis. CBR seems to be a suitable technique for medical knowledge based application. This technique will be more effective at applying the existing cases to new situation. It will be as the doctor diagnostic assistant as the aim of the research is to help doctors classify the class of tumour as either *benign* or *malignant* group. In order to achieve the aim of the research, the objectives such as developing an intelligent decision support application, applying the CBR algorithm in the breast cancer diagnosis application must be met. The methodology of the application will be discussed in the next chapter**.**

CHAPTER III

METHODOLOGY

3.0 Introduction

This chapter will be discussing the methodology and the framework of my project, diagnosis of breast cancer using case-based reasoning. This application is called as Breast Cancer Diagnosis Application (BCDA). This project will be conducted based on the Rapid Application Development (RAD) which is a software development life cycle designed to give much faster development and higher quality results than the traditional life cycle. There are four phases in RAD such as requirement planning, user design, construction and cutover.

3.1 Rapid Application Development

RAD is one of the software development methodologies that still widely used today. It is a based upon a structured step-to-step approach to developing information applications that promises better and cheaper applications and more rapid deployment by having applications developers and end users work together jointly in real time to develop applications (Hoffer.et.al, 2008). This methodology is created to decrease the time needed to design and implement information applications. According to Sommerville (2007), RAD techniques evolved from so-called fourth-generation language in the 1980s and are used for developing applications that are data-intensive. The life cycle of RAD is shown in Figure 3.1 as follow:

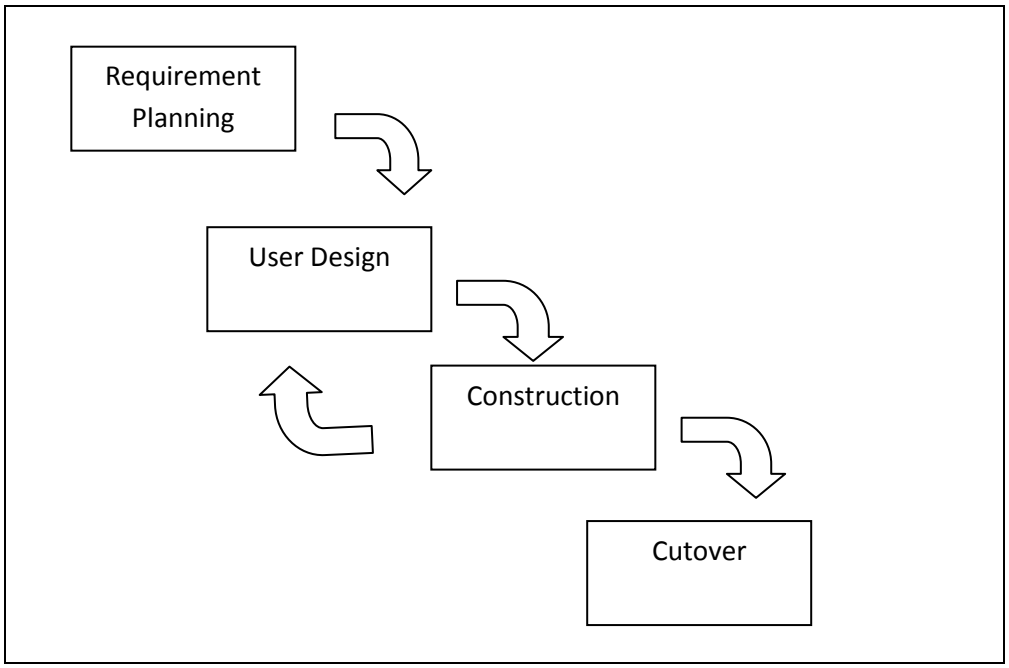


Figure 3.1: RAD Life Cycle

Martin (1991) suggested the life cycle of RAD consists of four phase, which are requirement planning, user design, construction and cutover.

3.1.1 Requirement Planning

Requirement planning is the first phase in RAD. The purpose for this phase is to determine the new user requirement. All the application objectives, application scope, application requirement is discussed and agreed. In this phase, all the data and information that used for developing the application are gathering include the data source and sample, software, hardware.

3.1.1.1 Data source and sample

In the project, sample of 100 data of the breast cancer Wisconsin dataset is selected and used. The sample was retrieved from UCI Machine Learning Repository. Each sample record has nine attributes and graded on an interval scale from 1 to 10. They are queued in order of ordinal data type (ordered set). The nine integer-valued attributes used are clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. Table 3.1 will show all the details of important features in diagnosis breast cancer.

Table 3.1: Nine important attributes in diagnosis breast cancer

Name of attributes	Type of value
Clump Thickness	Ordinal (integer value in range 1 to 10)
Uniformity of Cell Size	Ordinal (integer value in range 1 to 10)
Uniformity of Cell Shape	Ordinal (integer value in range 1 to 10)
Marginal Adhesion	Ordinal (integer value in range 1 to 10)
Single Epithelial Cell Size	Ordinal (integer value in range 1 to 10)
Bare Nuclei	Ordinal (integer value in range 1 to 10)
Bland Chromatin	Ordinal (integer value in range 1 to 10)
Normal Nucleoli	Ordinal (integer value in range 1 to 10)
Mitoses	Ordinal (integer value in range 1 to 10)

The class attribute is came with two states; 2 for *benign* and 4 for *malignant*.

3.1.1.2 Hardware

In this study, the application will be develop using the Intel (R) Celeron (R) processor 550 laptop with 1.5GB of RAM.

3.1.1.3 Software

Software that will use for developing this application is NetBeans IDE 6.9 for creating the interface and code the function of the application. The NetBeans IDE 6.9 used Java as the programming language.

3.1.2 User Design

User Design is the RAD phase during which the joint application design team examines requirements and transforms them into logical descriptions. It is a continuous interactive process that allows users to understand, modify, and eventually approve a working model of the application that meets their needs. The application design can be planned as a series of iterative steps or allowed to evolve.

In user design, the flowchart is used to describe the flow of BCDA. When the new breast cancer case is determined, the application is searched the case base for the similar and retrieved the most similar case from the case base. If the case base is matched with the new case, the application reused the solution from the similar case as the new solution. If no, the process is going back to the retrieved the most similar case. In this application, some part of CBR is used which are retrieve and reuse. The flow chart of breast cancer diagnosis is shown in Figure 2.

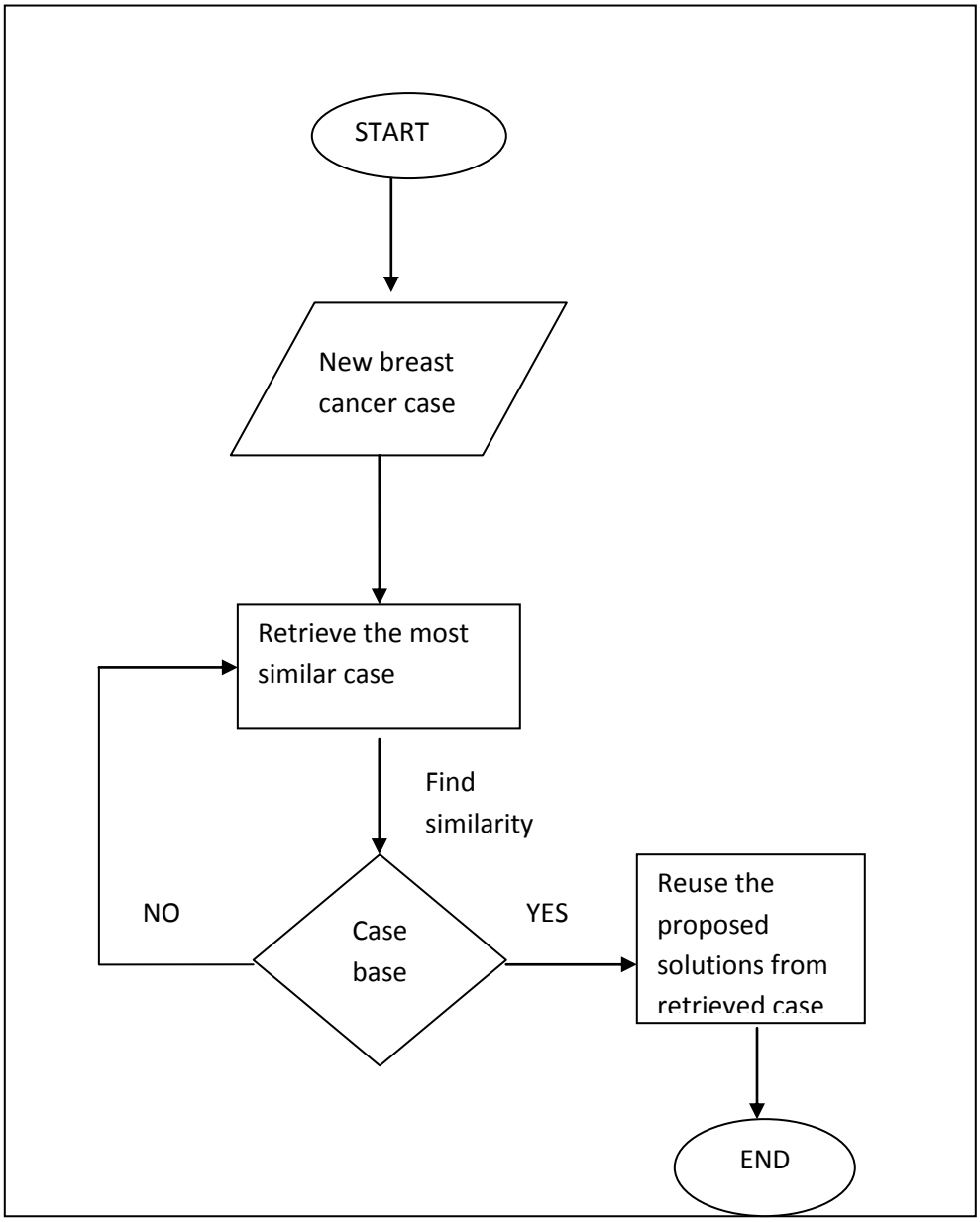


Figure 3.2: Flow chart of diagnosis of breast cancer

Figure 3 below shows the use case diagram of the application.

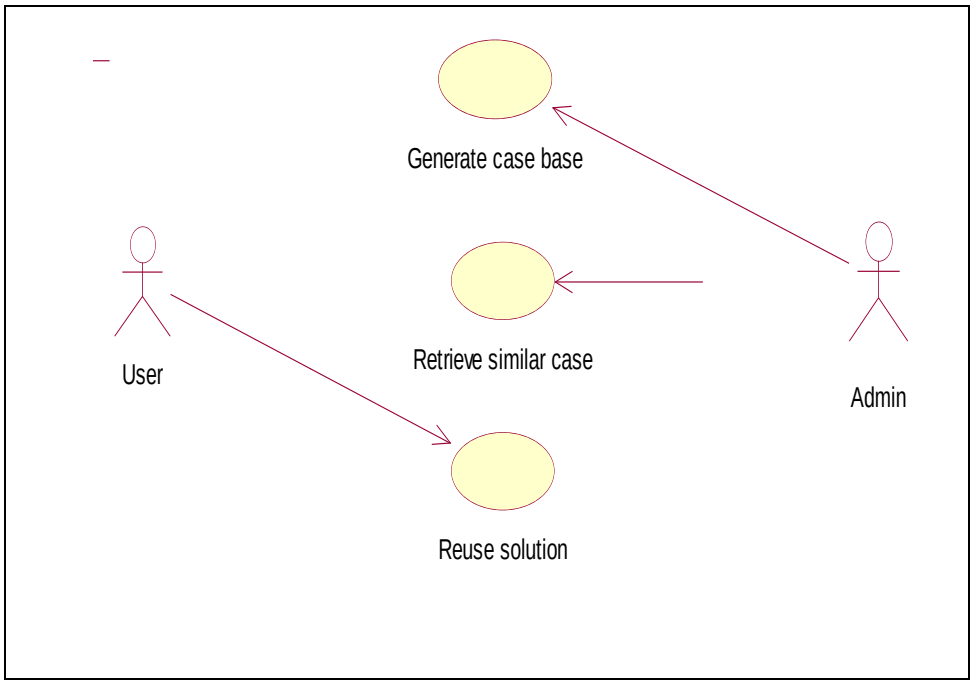


Figure 3.3: Use Case diagram

In Figure 3, the use case diagram is used to describe the actor and their roles in the application. The user is expert doctors or medical staff whose are responsible to do the diagnosis. When the new breast cancer case is determined, the admin generated the application to case base for search the similar case. Then the application retrieved the most similar case and match with the new case. If the similar case is matched with the new case, the expert doctors reused the proposed solution from the retrieved case as the solution of the new case.

3.1.3 Construction

In this phase, a prototype is built, exercised and modified based on user feedback. Its tasks are programming and application development, coding, unit-integration and application testing. The application developer start built the prototype for the interface.

This application is a stand-alone application. It is coding in Java language. In this application, some parts of CBR are used for the solution of new breast cancer case which is retrieve and reuse. The modification of the application is continued until the end, the acceptable and functional version is merged based on feedback of user. The implementation of the application will be discussed in detail in Chapter 4.

3.1.4 Cutover

Cutover phase is the last phase in RAD which the application is finalised and released to the user. This phase includes final user testing and training, data conversion and the implementation of the application.

After the BCDA is completely done, the application is tested again until it is achieve the objective of this application and can be used for the user in smoothly and fast. The application always be upgrade and tested to increase the user satisfaction and user-friendly.

3.2 Conclusion

RAD is an object-oriented approach to applications development that includes a method of development as well as software tools. It is depending on prototyping. RAD methodology is suitable for the BCDA because the user design and construction can be done rapidly until the successful version is emerged. CBR reuse the existing data that have been stored in case base as the solution for the new case that are similar. By the end of developing the application, hopefully it will functioning well in assign the patients to either a *benign* group that does not have breast cancer or a *malignant* group that has strong evidence of having breast cancer. The implementation of the application will be discussed in next chapter.

CHAPTER IV

IMPLEMENTATION

4.0 Introduction

This chapter will be discussed about how the implementation stage has been done in developing the BCDA. The development is involving the Graphical User Interface (GUI) design, database design and the coding development for entire application. The method used in developing the application and database also is discussed here.

BCDA has been developing using the NetBeans IDE 6.9. The source code of the application using the Java programming language and the database was created using the SQLite database.

4.1 Development Environment

For this application, it is developed in the NetBeans IDE 6.9 using the Java programming language. Window XP is used as the operating system with Intel(R) Celeron(R) processor 550 and 1.5G of RAM for develop the application environment. This application used the Wisconsin Breast Cancer dataset for evaluating the CBR algorithm. This dataset is retrieved from UCI Machine Learning and stored in SQLite database. The dataset is used as the database for the application. Table 4.1 shows the environmental needs for the application development.

Table 4.1: Environmental needs

Type	Tool	Platform
Programming Platform	Netbeans IDE 6.9	Windows
Programming Language	Java	Windows
Operating System	-	Window XP
Hardware	-	Acer Aspire 4315 laptop
Processor	-	Intel (R) Celeron (R) processor 550 with 2.0GHz
RAM	-	1.5GB
Database	SQLite Manager SQLite Database Browser 2.0 b1	-

To connect the SQLite database, the Java driver is needed. SQLiteJDBC is a Java driver for SQlite. It runs using either a native code library 100% Pure Java driver based on NestedVM emulation. Both the pure driver and the native for Windows, Mac OS X, and Linux x86 have been combined into a single jar file. Both of them have to be added into the Netbeans libraries and also rs2xml.

4.2 Designing of Interface

Interface is the layer of the application or system that used to interacts with others. It is used by the user to interact between one interface with another interface. For BCDA, it consists of one main interface with many panels inside the interface. The first panel is Diagnose interface, second panel is Result interface. There also interface briefly explain about the BCDA and breast cancer. The Database interface will display the database loaded and Exit interface allows user to exit the application.

4.2.1 Main Interface

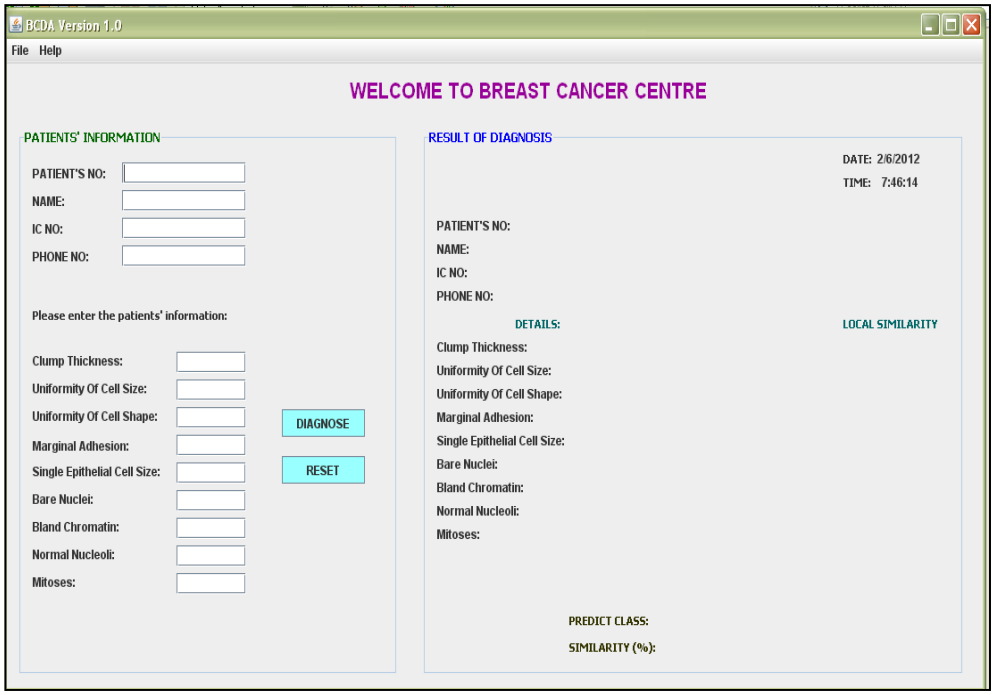


Figure 4.1: Main Interface (Menu) of BCDA

This is the main interface of the BCDA. It consists of three panels which are first panel is for Diagnose panel, second panel is for Result panel and third panel is for the buttons panel. Diagnose panel have the textfields that have to be filled in as the input for diagnosis. While in the Result panel will show the result of the diagnose.

4.2.1.1 Diagnose Panel

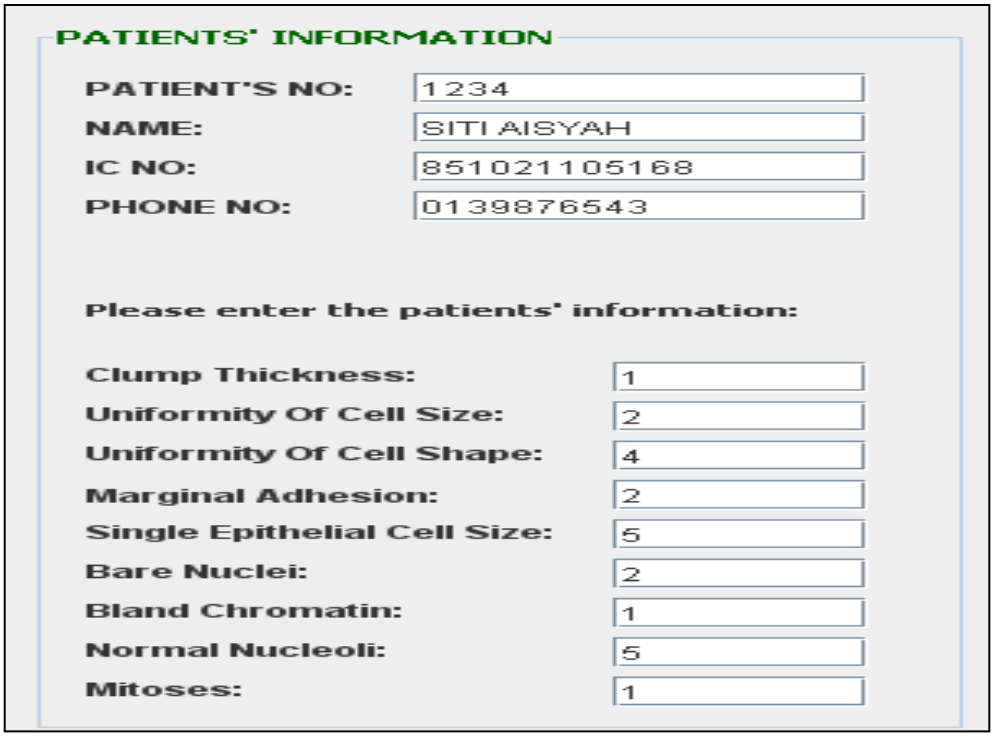


Figure 4.2: Diagnose Panel of BCDA

Diagnose panel is the page for the user to input the patients' information about their breast cancer. From the diagnose interface, after click the diagnose button, the input will connect with the database, retrieve the data from the database to do the calculation of local and global similarity of the diagnosis. After the calculation, the result will display the class of the breast cancer with the information of patient and the similarity from the previous case.

4.2.1.2 Result Panel

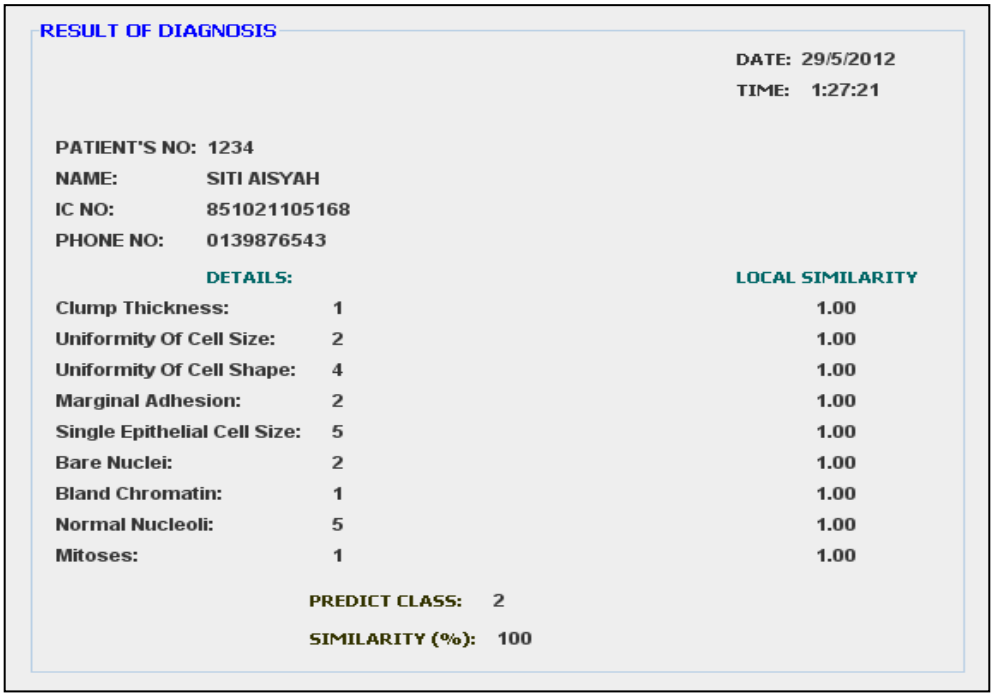
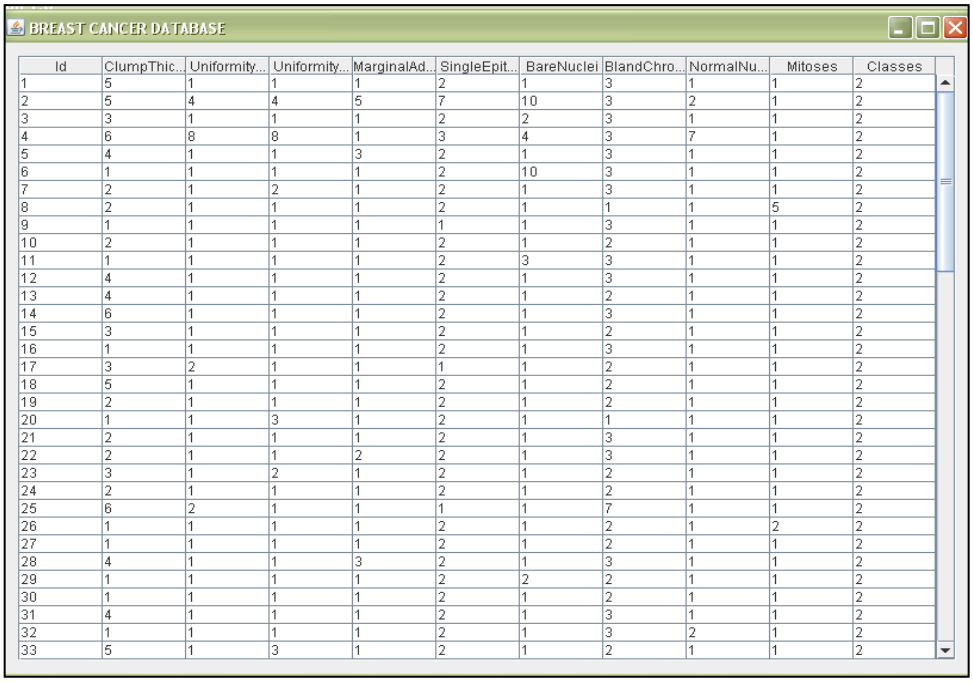


Figure 4.3: Result Panel of BCDA

This is the result interface where the result of the diagnosis will display. In the result interface, it will display the information of patient, similarity of the case to the previous case and the class of the case. User can choose either to back to diagnose page or to the end page which is exit interface.

4.2.2 Load Database Interface



Id	ClumpThic...	Uniformity...	Uniformity...	MarginalAd...	SingleEpit...	BareNuclei	BlandChro...	NormalNu...	Mitoses	Classes
1	5	1	1	1	2	1	3	1	1	2
2	5	4	4	5	7	10	3	2	1	2
3	3	1	1	1	2	2	3	1	1	2
4	6	8	8	1	3	4	3	7	1	2
5	4	1	1	3	2	1	3	1	1	2
6	1	1	1	1	2	10	3	1	1	2
7	2	1	2	1	2	1	3	1	1	2
8	2	1	1	1	2	1	1	1	5	2
9	1	1	1	1	1	1	3	1	1	2
10	2	1	1	1	2	1	2	1	1	2
11	1	1	1	1	2	3	3	1	1	2
12	4	1	1	1	2	1	3	1	1	2
13	4	1	1	1	2	1	2	1	1	2
14	6	1	1	1	2	1	3	1	1	2
15	3	1	1	1	2	1	2	1	1	2
16	1	1	1	1	2	1	3	1	1	2
17	3	2	1	1	1	1	2	1	1	2
18	5	1	1	1	2	1	2	1	1	2
19	2	1	1	1	2	1	2	1	1	2
20	1	1	3	1	2	1	1	1	1	2
21	2	1	1	1	2	1	3	1	1	2
22	2	1	1	2	2	1	3	1	1	2
23	3	1	2	1	2	1	2	1	1	2
24	2	1	1	1	2	1	2	1	1	2
25	6	2	1	1	1	1	7	1	1	2
26	1	1	1	1	2	1	2	1	2	2
27	1	1	1	1	2	1	2	1	1	2
28	4	1	1	3	2	1	3	1	1	2
29	1	1	1	1	2	2	2	1	1	2
30	1	1	1	1	2	1	2	1	1	2
31	4	1	1	1	2	1	3	1	1	2
32	1	1	1	1	2	1	3	2	1	2
33	5	1	3	1	2	1	2	1	1	2

Figure 4.4: Load Database Interface of BCDA

This page is called as Load Database. It contains the data that retrieve from the SQLite Manager. It is displayed in Grid View and just the view page to view the database. The database is stored in SQLite Manager and can be browser using SQLite Database Browser 2.0 b1 and also named as dataset.sqlite.

4.2.3 Breast Cancer Info Interface

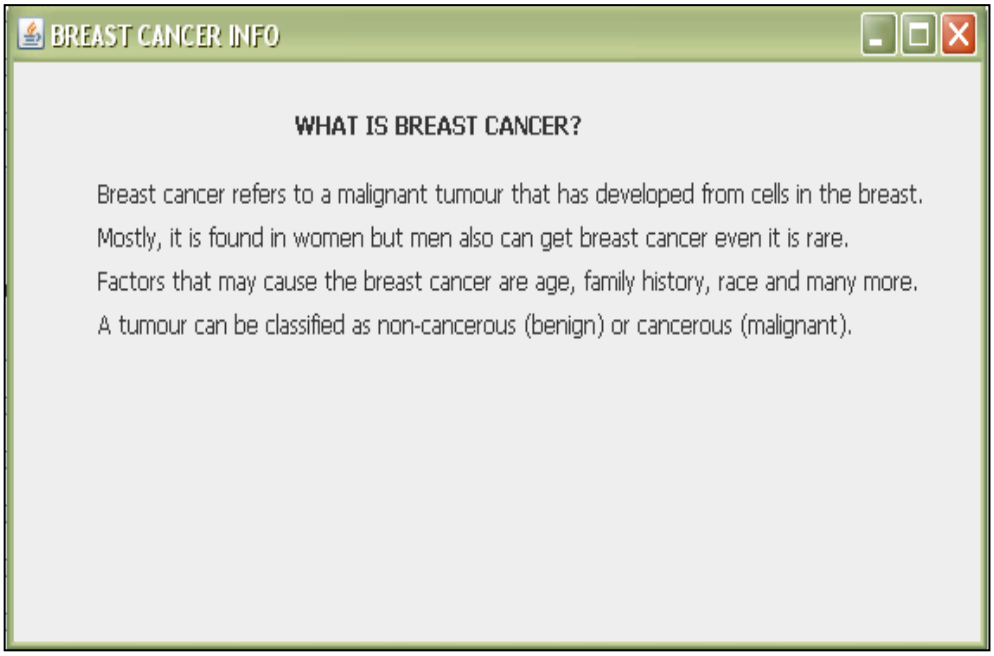


Figure 4.5: Breast Cancer Info Interface of BCDA

This page consists of briefly explanation about the breast cancer. it is for the general view of user about the cancer.

4.2.4 BCDA Version 1 Interface

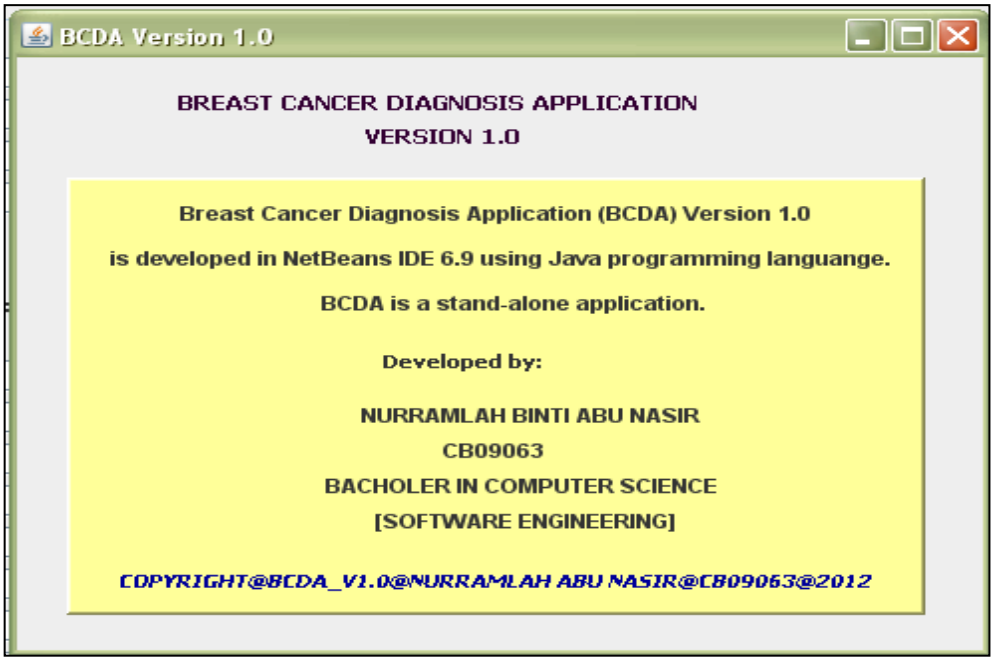


Figure 4.6: BCDA Version 1 Interface of BCDA

This page is about the BCDA. In this page, it is displayed the information about the developer of BCDA, platform of the application and also language used during developing the BCDA.

4.2.5 Exit Interface

Figure 4.7: Exit Interface of BCDA

Exit interface is the last page of the application. This page just shows appreciation to the user for using the application before they leave out the BCDA.

4.3 Database Design

In this section, the process of producing a detailed data model of a database is discussed. The database management and implementation also will be detailed here. The database is stored in SQLite Manager with sqlite file format. It can be browser using the SQLite Database Browser 2.0 b1.

4.3.1 Database in BCDA

Figure 4.8 shows the list of breast cancer database in SQLite Database Browser 2.0 b1. There is one table in the database which contain all the dataset from previous cases. This dataset called case-based data that contain nine attributes with which are Id, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nuclei and Mitoses. For the class attribute, there are two result of the class either Number 2 for *benign* group or Number 4 for *malignant* group.

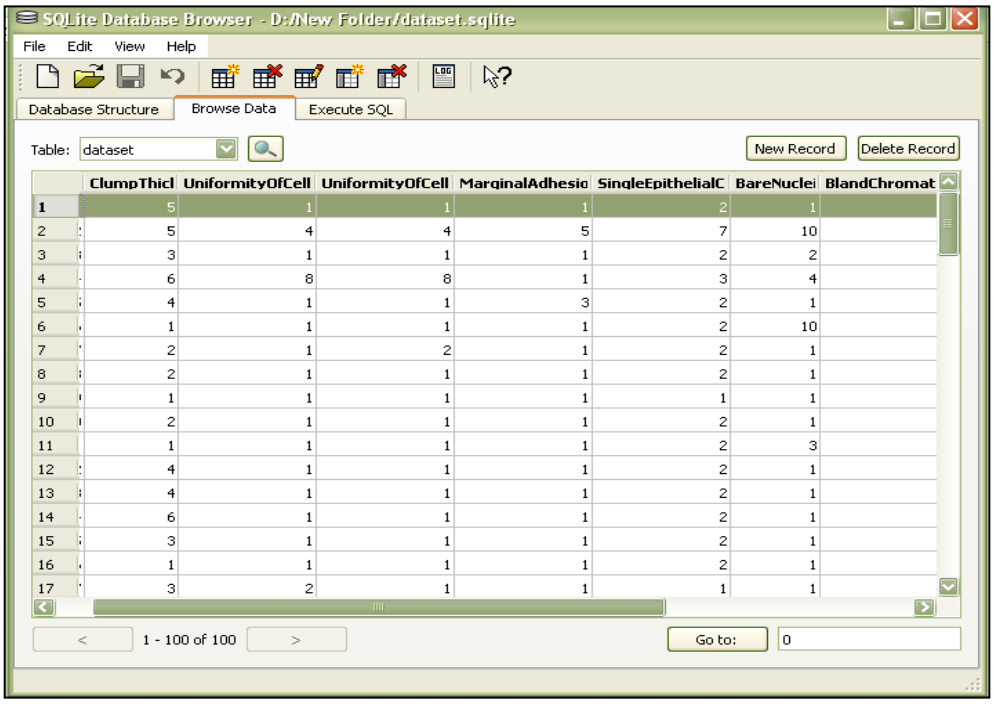


Figure 4.8: Breast Cancer Dataset Table is Database

The dataset of breast cancer is saved as dataset.sqlite. This table contain of 100 data of breast cancer from UCI Machine Learning.

4.3.1.1 Database Connection

Connection with database is the most important part in developing this application. The connection must be done before applying the CBR algorithm. It is important to connect with the database because the new cases are solved by recalling from previous solved problem which are

store the case-base. The code on how to create the connection of database is shown at Appendix F.

4.4 Breast Cancer Diagnosis Application Engine Module

BCDA Engine Module consists of two main components which are training data and testing data. Testing data will used the new data which indicate as the new case component. Training data is used to measure the accuracy of data while testing data is used to measure the accuracy of algorithm.

In CBR, there are four components that are important during the prediction which are Retrieve, Reuse, Revise and Retain. In developing of the BCDA, it just used two out of four components of CBR which are focus on Retrieve and Reuse only. Retrieve is referring to given a target problem, retrieve cases from memory that are relevant to solve it. A case consists of a problem, its solution and about how the solution is derived. In BCDA, case retrieval refers to process of finding the nearest case, which includes the solution for the new case within the case-base. After the nearest case is retrieve, the solution from the previous case is reused to solve the new case.

4.4.1 Similarity Measure

Similarity measure is used in problem solving and reasoning to match a previous case (case-base) with the new case to find solution. It select cases that have nearly the same solution that the new case. There are two types of similarity which are Local Similarity and Global Similarity.

1) Local Similarity

Local similarity is a similarity between two cases. It is used to compute the similarity between query (new problem) and case attributes values. The similarity between two cases is based on the local similarity of each attributes of the case. The formula to calculate the local similarity will be shown in Equation 4.1 as below:

$$sim(a,b) = 1 - \frac{|a - b|}{range} \quad (4.1)$$

Where,

a = new features of the problem case

b = previous features of the retrieved case

range = the value of difference between the upper and lower boundary of the set

2) Global Similarity

After a set of local similarities have been calculated for each feature in the case, a global similarity of the case is calculated. Global similarity provides a case-matching behaviour using the global similarity calculation to find the relationship between two cases. The formula used to calculate the global similarity is shown in Equation 4.2 as below:

$$SIM(A, B) = \sum_{p} \frac{w_i sim_i(a_i, b_i)}{} * 100 \quad (4.2)$$

Where,

A = new case

B = previous case

a_i = new feature from local similarity

b_i = previous features from local similarity

p = Number of attributes

w*i* = Weight of attributes *i*

sim_i = local similarity calculated for attribute *i*

SIM = Global similarity for the case

4.5 Summary

In the nutshell, the BCDA design is presented with the development environment, designing of interface is discussed. The application applies the CBR technique which are retrieve and reuse. The next chapter will discuss about the testing and discussion about the result of the testing.

CHAPTER V

TESTING AND RESULTS

This chapter briefly explain about testing technique used, the result of the testing and following by the discussion of the project Diagnosis of Breast Cancer using Case-Based Reasoning (CBR).

5.1 Testing and result

The nine integer-valued attributes used are clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses as the input of the system. Similarity measure is used in problem solving and reasoning to match a previous case of breast cancer with the new problem to find solution. After find the similarity, the local and global similarities are calculated. A local similarity is calculated between query (new problem) and case attributes values. After a set of local similarities have been calculated

for each feature in the case, a global similarity will be calculated. Global similarity provides a case-matching behaviour using the global similarity calculation to find the relationship between two cases. The most important features (weight) are determined and it will be used in similarity computation by weighted average. For estimating the accuracy rate of the CBR model, dataset is divided into two sets. One of them is training set that is used for model training and another is test set that is used for estimating accuracy of the model. So, 100 of data are allocated to training data and the 50 is allocated to testing data. The output of the system is the class attributes either *benign* or *malignant* group.

Decision support system has very clear target and the target will affect by algorithm and used database. CBR algorithm in this application will bring more than 80% accuracy of diagnose of cancer. Generally, artificial intelligent technique has different advantages and disadvantages. CBR would lead to 20% error of determination in limited amount of database. Once the target get expected result that same with the testing case, the case will be consider as true and higher the accuracy of system. The measure accuracy is computed in Equation 5.1 as follow:

$$\text{Accuracy} = \frac{\text{Correct Classified}}{\text{Total Testing Cases}} * 100 \qquad (5.1)$$

While, the percentage of error is computed as below in Equation 5.2:

$$\text{Percentage of error (\%)} = \frac{\text{Predict result} - \text{Actual result}}{\text{Actual result}} * 100 \qquad (5.2)$$

By using the testing data, from the testing result, 4% of errors happen while testing the 50 of different data from the same database. Out of 3% of errors is came from Class 2 (benign), while 1% of error is based on Class 4 (malignant). The percentage of accuracy is 96%. This means 46 cases out of 50 cases are similar to the original class.

While using the training set, application reached 98% accuracy which are 98 cases have same prediction result with the original result when predicting the result with database itself.

Table 5.1 show the data testing result for class 2 which is benign while Table 5.2 shows the data testing of class 4 which is malignant. The full result of the testing is shown in the Appendix D and E.

Table 5.1: The testing result for class 2 (Benign)

ID	CLASS	SIMILARITY
48	2	100.00
20	2	100.00
49	2	98.77
44	2	98.77
23	2	98.77

Table 5.2: The testing result for class 4 (Malignant)

ID	CLASS	SIMILARITY
77	4	90.12
62	4	90.12
57	4	79.01
95	4	76.54
54	4	75.31

CHAPTER VI

CONCLUSION AND RECOMMENDATION

6.0 Conclusion

The aim for this application is to help the expert doctors or medical staffs doing their breast cancer diagnosis. The purpose of this algorithm is to serve as doctor diagnostic assistant and aid the young physicians to check their diagnosis.

As mentioned in Chapter 1 Introduction, The objectives for this application are:

1) To develop an intelligent decision support application for diagnosis the breast cancer in order to classify the class of tumour either *benign* or *malignant* group.

2) To apply the CBR algorithm in the breast cancer diagnosis application.

The objectives of this application that stated are achieved. The result of the testing does not achieved 100%. It may due on the calculation and weight. To get more accurate result, the calculation of the similarity has to modify with the weight. The opinion and view from the expertises will determined the most important features (weight) and it will be used in similarity computation by weighted average. By doing that, the most accurate calculation will be determined.

The methodology used in the application is RAD because it promotes the accuracy application development and delivery and reduced the cycle time. CBR seems to be a suitable technique for medical knowledge based application. It can improve accuracy and problem solving performance through the reuse of the previous similar situation and knowledge that accelerated the application development process in diagnosis breast cancer. The classification of class of tumours either *benign* or *malignant* group can be done using this algorithm and this algorithm has a great potential to be implementation in diagnosis of breast cancer.

6.1 Recommendation

There are some suggestions and recommendations that should be done in order to improve the application as follow:

- Doing more research about the application for doing the improvements in the future for produce more quality application.
- Meet the expertises in the medical domain about the breast cancer to find out the most important attributes that they used in doing the diagnosis of breast cancer
- For next version this application can be implementing inside the mobile application due to the development of technology nowadays.

Finally, hopefully the development of the BCDA can motivate and inspire other people, researcher, lecturers or academia to continue and upgrading the development of the application to become more faster and can be used in large scale of medical domain.

REFERENCES

Aamodt, A., Plaza, E., (1994). *Case based reasoning: Foundation issues, methodological variations, and application approaches*. AI Communications. IOC Press, 7(1), 39-59.

ai-cbr.org (2011). Retrieved on October 31, 2011 from

http://www.ai-cbr.org/applied.html

cancer.org (2011). Retrieved on November 1, 2011 from

http://www.cancer.org. *Breast Cancer Facts & Figures 2009-2010*. American Cancer Society 2009. Atlanta: American Cancer Society, Inc.

Cortes, C. and Vapnik, V. N. (1995). *Support vector networks*, Machine learning Boston, vol.3, Pg.273-297.

Deng P.S. (1996). *Using Case-Based Reasoning Approach to the Support of Ill-structured Decisions,* European Journal of Operational Research 93, 511- 521.

Elsayad, A.M. (2010). *Diagnosis of Breast Tumor using Boosted Decision Trees*. ICGST-AIML Journal, 10 (1).

Fogel, D. B., Wasson III.E.C. and Boughton, E.M. (1995). *Evolving Neural Networks for Detecting Breast Cancer*. Cancer Letters 96 (1995) 49-53.

Gierl, L. and Stengel-Rutkowski, S. (1994). *Integrating Consultation and Semi-automatic Knowledge Acquisition in a Prototype-based Architecture: Experiences with Dysmorphic Syndromes*. ArtificIntell in Med 6, 29- 49.

Hoffer, J.A., George, J.F. and Valacich, J.S. (2008). *Modern Applications Analysis and Design*.Pearson Education, Inc. Fifth edition. 21.

Imaginis.com (2011). Retrieved on October 29, 2011 from http://www.imaginis.com/breasthealth/breast_cancer.asp

Investopedia.com (2011). Retrieved on October 31, 2011 from http://www.investopedia.com/terms/n/neralnetwork.asp

Investopedia.com (2011). Retrieved on October 31, 2011 from http://www.investopedia.com/terms/f/fuzzy-logic.asp

Janghel, R.R , Shukla, A., Tiwari, R. and Kala, R (2010). *Breast Cancer Diagnosis using Artificial Neural Network Models*. IEEE paper.

Kolodner, J. L. (1993). *Case-based Reasoning*. ISBN: 1-55860-237-2, Morgan Kaufmann, San Mateo.

Kovalerchuc, B., Triantaphyllou, E., Ruiz, J.F. and Clayton, J. (1997). *Fuzzy Logic in Computer-Aided Breast Cancer Diagnosis: Analysis of Lobulation.* Artificial Intelligence in Medicine 11 : 75-85.

Krovvidy, S. and Wee, W.G (1993). *Wastewater Treatment Applications from Case Based Reasoning. Machine Learning, 10, 341-363*. Kluwer Academic Publisher, Boston.

Lancaster, J.S and Kolodner, J.L, (1988). *Varieties of Learning from Problem Solving experience*. Proceedings of the Tenth Annual Conference of the Cognitive Science Society.

Mangasarian, O. L. and Wolberg, W. H. (1990). *Cancer diagnosis via linear programming*, SIAM News, Volume 23, Number 5, pp 1 & 18.

Martin, J. (1991). *Rapid Application Development*, Macmillan, New York.

Nilsson, M., Sollenborn, M., (2004). *Advancements and trends in medical case based reasoning: An overview of applications and application development*.In Proceedings of the 17th International FIAIRS Conference, Special Track on Case-Based Reasoning. American Association for Artificial Intelligence, Miami, USA, 178-183.

Pal, S.K. and Shiu, S.C.K. (2004). *Foundations of soft case-based reasoning*. Wiley- Interscience. John Wiley & Sons, Inc. Publication. Hoboken, New Jersey. 10-11.

Pantazi, S.V., Arocha, J.f. and Moehr, J.R. (2004). *Case-based Medical Informatics*. BMC Medical Informatics and Decision Making 2004, 4:19. doi: 10.1186/1472-6947-4-19

Roukis, G.S., Conway, H. and Charnov, B.H. (1990). *Global Corperate Intelligence: Opportunities, Technologies and Threats in the 1990s*. Greenwood Publishing Group 1990 – Business and Economies. Page 68.

Schmidt G. (1998). *Case Based Reasoning for Production Scheduling*, International Journal of Production Economics 56-57, 537-546.

Sommerville, I. (2007). *Software Engineering*. Pearson Education Limited. Eighth edition.405-409.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag.

WangLing(2001). *Intelligent optimization algorithms and its application*. Beijing: Tsinghua University press.

Welsh, D.J. A. and Powell, M. B. (1967). *An Upper Bound on the Chromatic Number of a Graph and its Application to Timetabling Problem*. The Computer Journal 10, 85-86.

Wikipedia.org (2011). Retrieved on October 31, 2011 from http://en.wikipedia.org/wiki/Clinical_decision_support_application#Genetic Algorithms